Task A: Review a related paper (100 points)

Review the paper – Women through the glass ceiling: gender asymmetries in Wikipedia
https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0066-4

Please read the paper carefully and write 1-2 paragraphs each to summarize the following in your words:

- **Primary goal of the paper (30 points)**
- **Main methods (20 points)**
- **Key findings (30 points)**
- **Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and compute resources (20 points)**

1. Primary goal of the paper (30 points)

The first primary goal of this paper is to investigate the possible underlying gender and linguistic bias against women on Wikipedia pages, a global encyclopedia on various topics. Particularly, this paper's research is limited to personal biographies of men and women in the English language editions because this language has the most pages, and other languages have nearly 100% overlapping pages with English. The paper claims that the editor community has a narrow diversity, with a majority of white male editors, so there should exist a potential bias in the content of the biographies. As a result, the paper tries to investigate the properties of Wikipedia articles: notability, topical focus, linguistic bias, structural properties, and meta-data presentation. Based on these aspects, they can compare if articles about men and women have nearly similar structures or whether there is an inherent difference based on gender only.

The paper proposes a concept of the glass-ceiling effect, which refers to the situation in which women cannot reach higher positions due to an invisible barrier of gender bias. It hypothesizes that if the entry point of Wikipedia functions as a glass ceiling, fewer women will be included and those included will be more notable than their male counterparts on average. Based on its findings about the glass ceiling effect and biased linguistic uses against women, this paper concludes that there is indeed a gender inequality in Wikipedia articles. As such, the second primary goal of the paper is to promote Wikipedia's internal policies for better gender equality representation. The paper expects its work can contribute to increased awareness about gender biases online. It also suggests that Wikipedia editors consider revising its guidelines to account for women's low visibility and encourage a less biased use of language.

2. Main methods (20 points)

To study gender bias in Wikipedia, the authors use The DBpedia dataset and split it into pre-1900 and post-1900. The pre-1900 sample contains all people born before 1900, while the post-1900 sample consists of people born in or after 1900. The main methods this paper uses to assess the extent to which gender bias manifests in Wikipedia are the comparisons between articles about men and women along the following dimensions: global notability according to external and internal proxy measures, topical focus, linguistic bias, structural properties like meta-data and hyperlink networks. Furthermore, these comparisons are applied for both pre and post-1900 datasets since the authors believe the advent of the Internet significantly affects the varying level of bias in Wikipedia articles.

In detail, this paper uses the number of language editions as an internal measure (Wikipedian editors) and Google search volume as an external measure to quantify public interest. For topical bias, the paper analyzes three topics that could be over-represented in articles about women: gender, relationship, and family. For linguistic bias, the paper uses a lexicon-based approach and syntactic annotations to detect abstract and subjective language. For structural properties, this paper relies on the search result rankings, which are often informed by centrality measures such as PageRank. For meta-data, it compares the relative proportions of attribute presence between genders using chi-square tests. Finally, the paper computes the PageRank of articles about people to estimate the hyperlink networks.

3. Key findings (30 points)

The findings this paper claims are based on the proposed speculations in its methodologies. Regarding inequalities in global notability thresholds, this paper found that the gap between men and women is larger for people with low or medium levels of global notability than for global stars. This paper explains the high men-to-women ratio for local heroes: men love to create articles about themselves because they are more self-absorbed than women, or more information may be available online about less notable men than about them less notable women. However, this paper is hesitant to claim the glass ceiling effect for inclusion in Wikipedia of women born before 1900 because the more historic a person is, the more notable they are globally. Conversely, young people born recently may benefit from the availability of digital information, making them more likely to be recognized by editors. Another critical finding about Google search trends is that women in Wikipedia are slightly more of interest to the world.

Regarding topical and linguistic asymmetries, this paper finds that besides professional and topical areas, words in the gender, relationship, and family categories are more dominant in articles about women born before 1900 but less pronounced in articles about people born since 1900. Overall, women tend to have more words related to family, gender, and relationships than men. Additionally, adjectives are much more likely to describe the positive aspects of men's biographies and the negative aspects of women's biographies. Regarding structural inequalities, many attributes such as activeYearsEndDate, careerStation, and position are more frequently used to describe men. All of these attributes are related to sports due to the prominence of men in sports-related. Attributes deathDate, and deathYear are more frequently used for men born before 1900 because women's life was less well documented back then. In contrast, attribute birthName and spouse is more frequently used for women in recent times because married women change their surnames to those of their husbands in some cultures. Furthermore, women's biographies have more links to other women's articles, which stems from the reported interests of female editors in women's biographies in Wikipedia. However, a final key finding of this paper claims that the empirically observed structure of the hyperlink network puts women disadvantage and that there exists a bias in the generation of links by Wikipedia editors, favoring articles about men

4. Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and computing resources (20 points)

The paper mentions that their empirical results are limited to the English Wikipedia, which is biased towards western cultures. They wish to leave a more detailed exploration of gender bias across all language editions for future work. Therefore, one possible follow-up analysis that could be done based on the paper if unlimited access to data and computing resources were available is to expand the study to multiple language editions of Wikipedia or even other platforms to examine if the observed gender inequalities are consistent across different cultures and languages. This may involve applying the same methodologies in this paper to other language editions and comparing the results to those obtained in the English edition on Wikipedia. Because English puts more emphasis on the Western cultural sphere, this research results may not be true for the countries in the Eastern sphere, such as China, India, or Japan, where gender bias is harder to be quantified. Such analysis could provide insights into the extent to which gender disparities in biographies are a global phenomenon or only specific to the Western world.

For example, China is a highly secluded country from the world's digital network. Only highly notable individuals from mainland China are mentioned on Wikipedia, leaving possibly hundred thousand notable people inside China undocumented on Wikipedia. Given unlimited access to data, we can try to access Chinese databases to see whether there exists gender bias in Chinese media (although, in reality, this could be impossible). Additionally, given unlimited computing resources, we could identify language-specific issues that affect gender inequalities in each Wikipedia language edition and inform specific strategic developments to narrow the gender imbalance gap. This could provide valuable insights into the global extent of gender disparities on Wikipedia. The restricted confidentiality of data and limited resources for conducting analysis are the main constraints on this task. However, it is valuable research given the hypothesis that we have unlimited access to data and computing resources.